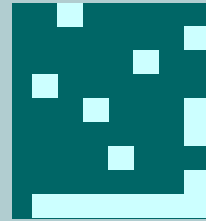


Data Editing and Imputation

This section introduces the editing and imputation procedures applied to SIPP data.

- *Types of Missing Data*
- *Problems with Missing Data*
- *Handling Missing Data*
- *Goals of Data Editing and Imputation*
- *Effects on Variance Estimation*
- *Processing SIPP Data*
- *Confidentiality Procedures*



Types of Missing Data

In SIPP, as in all surveys, both unit and item nonresponse may occur:

- Unit nonresponse occurs in SIPP when one or more of the people residing at a sample address are not interviewed and no proxy interview is obtained.
- Item nonresponse occurs when a respondent completes most of the questionnaire but does not answer one or more individual questions.



Problems with Missing Data

Missing data cause a number of problems:

- Analyses of data sets with missing data are more problematic than analyses of complete data sets.
- Analyses may be inconsistent because analysts compensate for missing data in different ways and their analyses may be based on different subsets of data.
- In the presence of nonresponse that is unlikely to be completely random, estimates of population parameters may be biased.

Handling Missing Data

The Census Bureau uses three different approaches for handling missing data in SIPP:

- Weighting adjustments are used for most types of unit nonresponse.
- Data editing (also referred to as logical imputation) is used for some types of item nonresponse.
- Statistical (or stochastic) imputation is used for some types of unit nonresponse and some types of item nonresponse.

Weighting is discussed in the Sampling Weights section of the tutorial (as well as in Chapter 8 and Appendix C of the *SIPP Users' Guide*).

Goals of Data Editing and Imputation

Data editing is the preferred method of handling missing data, and it is used whenever a missing item can be logically inferred from other data that have been provided. For example, when information exists on the same record from which missing information can be logically inferred, Census staff use that data to replace the missing information.

Analyses of survey data are usually based on assumptions about patterns of missing data. When missing data are not imputed or otherwise accounted for in the model being estimated, the implicit assumption is that data are missing at random after the analyst has controlled other variables in the model.

In SIPP, imputation procedures are based on the assumption that data are missing at random within subgroups of the population.

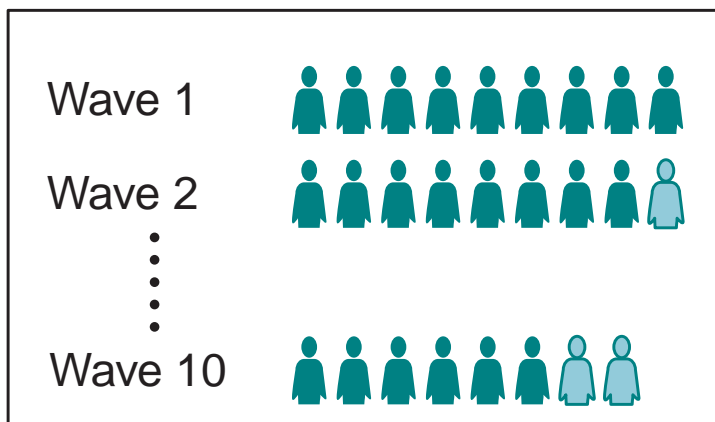
The statistical goal of imputation is to reduce the bias of survey estimates. This goal is achieved to the extent that systematic patterns of nonresponse are correctly identified and modeled. Unlike data editing, imputation results in an increase in variance.

The Census Bureau has been improving SIPP imputation procedures continually. With the 1996 redesign, the processing procedures for the wave files were enhanced with methods that use prior wave information to inform the editing and imputation of current waves (see Chapter 4 of the *SIPP Users' Guide*).



Effects of Imputed Data on Variance Estimation

Imputation fills in gaps in the data set and facilitates analyses. It also allows more people to be retained as panel members for longitudinal analyses. However, imputation changes data to some degree, and treating imputed values as actual values may lead to overstatements of the precision of the estimates. It is important that analysts recognize this fact when sizable proportions of values are imputed.



Processing SIPP Data

SIPP data are processed in two phases:

Phase 1: At the conclusion of each wave of interviewing, the Census Bureau processes the core and topical module data collected during that wave and creates the core wave and topical module files.

Phase 2: At the conclusion of the final wave of interviews in a panel, the Census Bureau links core data from all waves and applies a new set of edit and imputation procedures to create the resulting full panel file.

Phase 1 Summary. During the first phase of SIPP data processing, the Census Bureau performs the following six tasks.

1. As each wave of interviewing is completed, core data collected during the wave are edited for internal consistency.
2. Following data editing, Census staff use statistical matching and hot-deck procedures to impute missing data from the core wave file. (See Chapter 4 of the *SIPP Users' Guide* for a description of the imputation procedures.) *tip*

SIPP *tip*

Imputation can introduce inconsistencies into the data. When users detect inconsistencies, they should check the allocation (imputation) flag to see if the inconsistent data might have been imputed. See Chapter 4 of the SIPP Users' Guide for more information.

3. Census staff then create a public use version of the core wave file from the internal core wave file. They suppress or topcode selected information in the public use file to protect the confidentiality of survey respondents.
4. On a separate production track from the core data, Census staff edit data from the topical module administered with the wave for internal consistency. The extent of data editing varies across the topical modules, and some topical modules receive almost no editing.
5. Next, staff members use hot-deck procedures to impute missing data in the topical modules. Again, the extent of imputation varies across the topical modules; some topical modules have no missing data imputed.
6. Census staff then create a public use version of the topical module file. They suppress selected information in the public use file to protect the confidentiality of survey respondents.

These six tasks are repeated at the end of each wave of interviews. Prior to the 1996 Panel, each wave was processed independently of other waves of data. Thus, when multiple core wave files are linked, apparent changes in a respondent's status could be due to different applications of data edits and imputations to the files being combined.

With the 1996 data, the hot-deck procedure was redesigned to rely on historical information reported in prior waves. In addition, other forms of longitudinal imputation, such as carry-over methods, were adapted.

Phase 2 Summary. At the conclusion of each panel, the Census Bureau creates a full panel file containing core data from all waves. Four steps are involved.



1. Core data from all waves are linked. Those data have already been subjected to the Phase 1 edit and imputation procedures.
2. Census staff apply a series of longitudinal edits to the full panel file. Unlike the core wave edit procedures, these edits are designed to create longitudinally consistent records for each person. Both reported values and values that were imputed during the first phase of processing are subject to change. Thus, the data in a full panel file may differ from the data in the core wave files from which the full panel file was constructed.
3. A missing wave imputation procedure is then applied. Data are imputed when a sample member was absent for one or two consecutive waves but was present for the two adjacent waves. Data for the missing wave(s) are interpolated on the basis of information from the fourth month of the prior wave and the first month of the subsequent wave. The missing wave imputation procedure was introduced with the 1991 Panel. Earlier panels were not subjected to this procedure.
4. Census staff create a public use version of the full panel file from the internal file. They suppress selected information to protect the confidentiality of survey respondents.

Confidentiality Procedures for the Public Use Files

The Census Bureau edits respondents' records to protect their confidentiality. Two procedures are used:

- Topcoding of selected variables (income, assets, and age)
- Suppression of geographic information

Addresses as well as states and metropolitan areas with populations of less than 250,000 are not identified. Also, specific nonmetropolitan areas (such as counties outside of metropolitan areas) are never identified.

In certain states, when the nonmetropolitan population is small enough to represent a disclosure risk, a fraction of that state's metropolitan sample is recoded to nonmetropolitan status. Thus, SIPP data cannot be used to estimate characteristics of the population residing outside metropolitan areas (see Chapter 10 of the *SIPP Users' Guide* for more details).

